

CLAIMS

1. A method in a computer system for transforming a document into a canonical representation, the document having a plurality of sentences, each sentence having a plurality of terms, comprising:

for each sentence,

parsing the sentence to generate a parse structure having a plurality of syntactic elements;

determining a set of meaningful terms of the sentence from these syntactic elements;

determining from the structure of the parse structure and the syntactic elements at least one grammatical role for each meaningful term; and

storing in an enhanced data representation data structure a representation of each meaningful term associated with its determined grammatical role, in a manner that indicates a grammatical relationship between a plurality of the meaningful terms.